DOI: 10.1534/genetics.104.037077

Effects of Natural Selection on Interpopulation Divergence at Polymorphic Sites in Human Protein-Coding Loci

Austin L. Hughes,*[,1] Bernice Packer,[†,‡] Robert Welch,[†,‡] Andrew W. Bergen,[‡] Stephen J. Chanock[‡,§] and Meredith Yeager[†,‡]

**Department of Biological Sciences, University of South Carolina, Columbia, South Carolina 29208, [†]Intramural Research Support Program, SAIC-Frederick, NCI-FCRDC, Frederick, Maryland 21702, [‡]Core Genotyping Facility, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892-4605 and [§]Section on Genomic Variation, Pediatric Oncology Branch, National Cancer Institute, National Institutes of Health, Gaithersburg, Maryland 20892-4605*

Manuscript received October 1, 2004
Accepted for publication January 3, 2005

ABSTRACT

To develop new strategies for searching for genetic associations with complex human diseases, we analyzed 2784 single-nucleotide polymorphisms (SNPs) in 396 protein-coding genes involved in biological processes relevant to cancer and other complex diseases, with respect to gene diversity within samples of individuals representing the three major historic human populations (African, European, and Asian) and with respect to interpopulation genetic distance. Reduced levels of both intrapopulation gene diversity and interpopulation genetic distance were seen in the case of SNPs located within the 5′-UTR and at nonsynonymous SNPs, causing radical changes to protein structure. Reduction of gene diversity at SNP loci in these categories was evidence of purifying selection acting at these sites, which in turn causes a reduction in interpopulation divergence. By contrast, a small number of SNP sites in these categories revealed unusually high genetic distances between the two most diverged populations (African and Asian); these loci may have historically been subject to divergent selection pressures.

IN the human genome, a large number of single-nucleotide polymorphisms (SNPs) are believed to be in protein-coding genes (estimated to be between 50,000 and 250,000) (Brookes 1999; Gray *et al.* 2000; Risch 2000; Schork *et al.* 2000; Chanock 2001; Lohnmueller *et al.* 2003). Such SNPs seem a promising source of candidate genes for understanding the genetic contribution to complex diseases such as cancer and heart disease, especially when the protein is known to play a role in biological processes relevant to the disease of interest. Because >7 million SNPs have been reported in the public database (dbSNP, build 120), it is desirable to develop methods of sifting through this information to find likely candidates for disease association (Sunyaev *et al.* 2001; Wang and Moult 2001; Ng and Henikoff 2002; Ramensky *et al.* 2002; Botstein and Risch 2003; Fleming *et al.* 2003; Freudenberg-Hua *et al.* 2003; Hughes *et al.* 2003; Stitzel *et al.* 2003).

Evidence of the action of purifying selection at a given locus is evidence that a deleterious allele is present (Zhao *et al.* 2003). Purifying selection is the form of natural selection that acts to eliminate selectively deleterious mutations. For example, purifying selection is expected to act against mutations that have deleterious effects on protein structure by causing changes to functionally important amino acid residues or on gene expression by altering regulation (Kimura and Ohta 1974; Nei 1987). Estimation of gene diversity (heterozygosity) at 1442 SNP sites in an ethnically diverse sample of humans revealed consistently reduced gene diversities at sites where SNPs caused amino acid changes, particularly those predicted to be disruptive to protein structure (Hughes *et al.* 2003). Since SNPs are almost always biallelic, a relatively low gene diversity at a given SNP site (*i.e.*, the site at which the polymorphism occurs, which may also be referred to as a "SNP locus"; see Hughes *et al.* 2003) is equivalent to a low allelic frequency of the less common of the two alleles. The reduction of gene diversity at these SNP sites, in comparison to SNPs in the same genes not affecting protein structure, is evidence that purifying selection has reduced the population allelic frequencies of deleterious SNP alleles. This in turn suggests that slightly deleterious mutations are widespread in the human population and that examination of the patterns of SNP diversity across protein-coding loci can help guide the search for disease-associated SNPs (Hughes *et al.* 2003).

In addition to its effects on allelic frequencies within populations, natural selection is expected to affect the extent of divergence in gene frequency between populations. Therefore, when samples are derived from historically distinct subpopulations, examination of the pattern

[1] *Corresponding author:* Department of Biological Sciences, University of South Carolina, Coker Life Sciences Bldg., 700 Sumter St., Columbia, SC 29208. E-mail: austin@biol.sc.edu

of interpopulation divergence at SNP sites potentially provides a novel source of information about past natural selection that can be useful in guiding the search for candidate loci in disease-association studies. Furthermore, interpopulation divergence may provide evidence not only of purifying selection but also of positive selection leading to allele frequency divergence between populations.

To exploit information on interpopulation divergence as a source of information regarding natural selection at SNP sites, we performed a comparison between human populations of allelic frequencies at 2784 SNPs from 396 protein-coding loci. These loci were chosen because they are involved in fundamental biological processes (including development, the cell cycle, and immunity) believed to be relevant to cancer and other complex human diseases (Packer *et al.* 2004). The SNPs were restricted to known genes and heavily biased toward exons, intron-exon borders, and regulatory regions within 5 kb of the start or end of the open reading frames. Because of the availability of a greater number of SNPs in the 5′-untranslated region (UTR) of genes than in our previous sample (Hughes *et al.* 2003), we were able to test for purifying selection on these sites, some of which may be expected to be subject to functional constraint due to their role in the regulation of gene expression.

MATERIALS AND METHODS

Samples: Allele frequency data for 2784 SNP sites were taken from the SNP500 database (Packer *et al.* 2004). This database is based on bidirectional sequence determination of each SNP in 102 unrelated individuals from the Coriell Institute for Medical Research (Camden, NJ; http://locus.umdnj.edu/nigms/). On the basis of self-described ethnicity, these individuals were assigned to four populations: 31 non-Hispanic Caucasians, 24 African/African-Americans, 24 of Pacific Rim heritage, and 23 Hispanic. Because the "Hispanic" classification does not correspond to a major geographically defined historical subdivision of the human population, it was not included in the analyses reported in this article. The other three groups are referred to in the following as, respectively, European, African, and Asian. SNPs for analysis were chosen within or close to genes, and the selection of genes and SNPs for analysis was drawn from publicly available databases. For details of sequencing methodology see Packer *et al.* (2004) and Hughes *et al.* (2003). Allelic frequency data for all SNPs are available at http://snp500cancer.nci.nih.gov.

Statistical analyses: Gene diversity (heterozygosity) at a SNP site was estimated by $1 - \Sigma_{i=1} x_i^2$, where n is the number of alleles and x_i is the population frequency of the ith allele (Nei 1987, p. 177). For a given biallelic SNP site, the allelic frequency divergence between each pair of subpopulations was computed using the formula

$$d = 1 - [(x_1 y_1)^{1/2} + (x_2 y_2)^{1/2}],$$

where x_1 and y_1 are the frequencies of the first allele in each of the two subpopulations, respectively, and x_2 and y_2 are the frequencies of the second allele in each of the two subpopulations, respectively. The average of d over all loci is the average genetic distance (D_A) (Nei 1987, p. 216).

TABLE 1

Average genetic distance (D_A ± SE) between human populations based on 2784 SNP loci

	European	Asian
African	0.0286 ± 0.0008	0.0314 ± 0.0009
European		0.0182 ± 0.0006

SNPs were classified with respect to their location and effect on protein function as follows: (1) SNPs located in the 5′-noncoding region outside the 5′-UTR, (2) SNPs located in the 5′-UTR, (3) SNPs in the 3′-noncoding region outside the 3′-UTR, (4) SNPS in the 3′-UTR, (5) SNPs in introns, (6) synonymous SNPs in exons, (7) nonsynonymous SNPs in exons causing a conservative amino acid change, and (8) nonsynonymous SNPs in exons causing a radical change. Radical nonsynonymous changes included a small number of cases ($N = 5$) where a SNP introduced a stop codon and a larger number of cases causing an amino acid replacement involved two amino acids with a pairwise stereochemical difference >3.0 according to Miyata *et al.*'s (1979) scale (based on amino acid residue volume and polarity). Otherwise, missense SNPs were categorized as conservative. Miyata *et al.*'s (1979) is one of many scales that measure chemical distance between amino acids, all of which are significantly correlated with one another. Use of other scales in preliminary analyses yielded similar results to those using Miyata *et al.*'s (1979) scale. Previous analyses of gene diversity at a smaller number of SNP sites from the same population showed a similar pattern in the case of nonsense SNPs and radical nonsynonymous SNPs (Hughes *et al.* 2003); therefore, the two were combined in a single category in the present analyses.

RESULTS

The average genetic distances between Africans and Europeans and between Africans and Asians were significantly greater than that between Europeans and Asians (paired *t*-test; $P < 0.001$), as is expected given the African origin of modern humans (Nei and Livshits 1989; Tishkoff and Verrelli 2003) (Table 1). The average genetic distance between Africans and Asians was significantly greater than that between Africans and Europeans (paired *t*-test; $P = 0.001$). Because the African and Asian populations showed the greatest pairwise genetic distance, we focused particular attention on the distribution of genetic distances between these populations for different functional categories of SNPs.

In the African-Asian comparison, striking differences were seen among functional SNP categories with respect to genetic distance (d) (Table 2). Radical nonsynonymous SNPs and SNPs in the 5′-UTR showed the lowest mean d values (Table 2). For every category except SNPs in the 5′-UTR, the mean d was significantly higher than that for radical nonsynonymous SNPs (Table 2). A similar pattern was seen in the case of nonparametric tests, in which all categories except 5′-UTR showed significant differences from radical nonsynonymous SNPs (Table 2). Although the median d for 3′-non-UTR and for con-

TABLE 2

Mean (±SE) and median genetic distance (*d*) values at SNP loci in different functional categories

		African *vs.* Asian		All between-population comparisons	
SNP category	N	Mean ± SE	Median	Mean ± SE	Median
Silent					
5′-noncoding					
UTR	65	0.0157 ± 0.0030	0.0105	0.0190 ± 0.0026	0.0108
Non-UTR	305	0.0353 ± 0.0028***	0.0168*****	0.0283 ± 0.0018***	0.0169******
3′-noncoding					
UTR	312	0.0341 ± 0.0031**	0.0109****	0.0275 ± 0.0018***	0.0140*****
Non-UTR	156	0.0270 ± 0.0033*	0.0105****	0.0230 ± 0.0021*	0.0139****
Intron	1101	0.0332 ± 0.0016***	0.0109****	0.0275 ± 0.0010***	0.0140*****
Synonymous in exon	373	0.0315 ± 0.0024**	0.0109****	0.0249 ± 0.0015**	0.0140****
Nonsynonymous					
Conservative	428	0.0276 ± 0.0020*	0.0105****	0.0241 ± 0.0013**	0.0140*****
Radical	44	0.0157 ± 0.0044	0.0105	0.0144 ± 0.0026	0.0073

Tests of the hypothesis that mean *d* equals that for radical nonsynonymous SNP loci (*t*-test): *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$. Tests of the hypothesis that median *d* equals that for radical nonsynonymous SNP loci (Mann-Whitney test): ****$P < 0.05$; *****$P < 0.01$; ******$P < 0.001$.

servative nonsynonymous SNPs was identical to that for radical nonsynonymous SNPs, the Mann-Whitney test (which is based on average ranks) showed a significant difference in location between the former two categories and radical nonsynonymous SNPs (Table 2).

Mean *d* between populations (*i.e.*, the mean of African-European, African-Asian, and European-Asian comparisons) likewise showed the lowest mean and median values in the case of radical nonsynonymous SNPs (Table 2). The values for the 5′-UTR were not significantly different from those for radical nonsynonymous SNPs, but all other categories showed significant differences from radical nonsynonymous SNPs in both parametric and nonparametric tests (Table 2).

As in our previous analyses (Hughes *et al.* 2003), we found differences among SNP categories with respect to gene diversity (heterozygosity) (Table 3). Mean gene diversity was lowest in the case of radical nonsynonymous SNPs, and mean gene diversities for all categories except the 5′-UTR differed significantly from that for radical nonsynonymous SNPs (Table 3). Median gene diversities showed an identical pattern (Table 3). We computed correlation coefficients, both the Pearson correlation coefficient and the nonparametric Spearman correlation coefficient, between *d* and gene diversity for each SNP category (Table 3). In most cases, the correlation coefficients were modest but statistically significant (Table 3).

Although the results showed overall lower interpopulation genetic distances for SNPs in the 5′-UTR and for radical nonsynonymous SNPs, certain SNPs did not follow this trend. Table 4 lists cases of SNPs in these two categories for which the genetic distance between African and Asian populations was unusually high [>2 standard deviations (SD) above the mean for the category]. Also included are cases of conservative nonsynonymous SNPs for which the genetic distance between African and Asian populations was >2 SD above the mean for that category (Table 4). In any of these cases it is possible that the high interpopulation distance is due to chance, whether random genetic drift or sampling error. On the other hand, it is possible that in at least some of these cases, differentiation between populations has been driven by natural selection.

DISCUSSION

In previous analyses with the same study populations, we found significantly reduced gene diversity in the case of radical nonsynonymous SNP loci in comparison with other SNP categories in the same genes (Hughes *et al.* 2003). Since the gene diversity at radical nonsynonymous SNPs was reduced in comparison to silent SNPs in the same genes, the likely cause of reduced gene diversity at the former is purifying selection (Hughes *et al.* 2003); that is, natural selection acting to remove deleterious mutations (Kimura and Ohta 1974; Nei 1987). Similar results were reported by Sunyaev *et al.* (2001), Freudenberg- Hua *et al.* (2003), and Zhao *et al.* (2003).

In this study, examining twice as many SNPs as in our previous study, we found evidence of purifying selection not only at radical nonsynonymous SNPs, but also at SNPs in the 5′-UTR. The evidence of purifying selection took the form not only of decreased gene diversity at these SNP sites but also of reduced interpopulation divergence. Overall, the interpopulation divergence was positively correlated with mean intrapopulation gene diversity. It is expected that, if purifying selection is

TABLE 3

Mean (±SE) and median gene diversity values and correlations between gene diversity and genetic distance at SNP loci in different functional categories

	Gene diversity		Correlation coefficient (P)	
SNP category	Mean ± SE	Median	Pearson	Spearman (rank)
Silent				
5′-noncoding				
UTR	0.205 ± 0.028	0.082	0.203 (NS)	0.365 (0.003)
Non-UTR	0.254 ± 0.014*	0.139****	0.377 (<0.001)	0.638 (<0.001)
3′-noncoding				
UTR	0.255 ± 0.014*	0.140****	0.345 (<0.001)	0.570 (<0.001)
Non-UTR	0.245 ± 0.020*	0.103***	0.330 (<0.001)	0.580 (<0.001)
Intron	0.265 ± 0.008**	0.169****	0.307 (<0.001)	0.508 (<0.001)
Synonymous in exon	0.274 ± 0.013**	0.180****	0.290 (<0.001)	0.458 (<0.001)
Nonsynonymous				
Conservative	0.224 ± 0.012*	0.089***	0.409 (<0.001)	0.551 (<0.001)
Radical	0.142 ± 0.028	0.064	0.358 (0.017)	0.674 (<0.001)
All sites	0.253 ± 0.005	0.132	0.331 (<0.001)	0.535 (<0.001)

Tests of the hypothesis that mean gene diversity equals that for radical nonsynonymous SNP loci (t-test): *$P < 0.01$; **$P < 0.001$. Tests of the hypothesis that median gene diversity equals that for radical nonsynonymous SNP loci (Mann-Whitney test): ***$P < 0.05$; ****$P < 0.01$.

acting similarly in two populations to reduce frequency of a deleterious allele, the reduction in frequency will be similar in both populations, causing similar allele frequencies in the two populations. As a result, the genetic distance between populations will be reduced at a locus subject to purifying selection in both populations in comparison to a locus that is subject to random genetic drift in both populations.

Thus, our results demonstrate the utility of examining interpopulation divergence as a simple strategy for identifying SNP sites subject to purifying selection that thus are candidates for complex disease associations. Note that, in the present case, the numbers of individuals examined in each population were modest. Thus, this approach holds promise for providing an efficient method for identifying SNPs subject to purifying selection when available resources limit the number of individuals that can be examined.

Moreover, the methods used here are more readily applicable to the study of SNP sites than are such widely used tests for natural selection as those of Tajima (1989) and Fu and Li (1993). The latter tests are more applicable to sequencing studies than to SNP studies because they examine all polymorphic sites across a gene. In the case of SNP studies certain polymorphic sites may be unknown and thus may not be ascertained. Lack of ascertainment of certain SNPs is not a problem for the present method because known SNPs are compared within other known SNPs in different functional categories ascertained for the same set of genes. Furthermore, the methods of Tajima (1989) and Fu and Li (1993) assume a random sample of the population and are very sensitive to violations of that assumption (Hammer *et al.* 2003). Indeed, constructing a random sample of sequences from the entire human species raises numerous practical difficulties. The present methods, on the other hand, because they depend on comparisons between SNPs of different categories across the same set of genes, do not depend on the assumption of a random sample.

The population frequencies of the lower-frequency allele at SNP sites in the categories (5′-UTR and radical nonsynonymous SNPs) that showed evidence of purifying selection were typically in the range of 1–10%. Wong *et al.* (2003) similarly reported numerous nonsynonymous SNPs with similar allelic frequencies in a sample of 114 human genes. These frequencies are at least an order of magnitude higher than those of human genes causing severe Mendelian diseases, such as cystic fibrosis or Huntington's chorea (McKusick and Francomano 1997). This in turn suggests that the selection coefficients at many of these SNP sites are smaller than those at loci associated with severe disease.

Evidence of such mild purifying selection can be used to identify candidate loci for association with complex diseases, since SNPs associated with complex diseases are expected to be slightly deleterious, unlike the highly deleterious genes associated with Mendelian diseases. Thus, if nonsynonymous SNPs or SNPs in the 5′-UTR show lower gene diversity than other SNPs in the same genes, these are potential candidates for complex disease-association studies (Hughes *et al.* 2003). In addition, the present results suggest that nonsynonymous SNPs or SNPs in the 5′-UTR that show low levels of interpopulation divergence are potential candidates for complex disease-association studies.

The protein-coding loci chosen for this study were

TABLE 4

5′-UTR and radical synonymous SNPs with unusually large (>2 standard deviations above mean *d* value for category) genetic distance between African and Asian populations

Gene	Function	Amino acid change[a]	dbSNP ID[b]	*d*
5′-UTR				
GH1	Growth hormone 1	—	Rs6175	0.087
IL8	Interleukin-8	—	Rs2227538	0.079
LIG4	Ligase IV, DNA/ATP-dependent	—	Rs1805388	0.066
		—	Rs4987182	0.076
PTGS2	Prostoglandin-endoperoxide synthase 1	—	Rs55276	0.110
Radical nonsynonymous				
ADH1B	Alcohol-dehydrogenase class I, β-polypeptide	R370C	Rs2066702	0.087
GNRH1	Gonadotropin-releasing hormone receptor	W16S	Rs6185	0.176
Other nonsynonymous				
ADH1B	Alcohol-dehydrogenase class I, β-polypeptide	R48H	Rs1229984	0.225
AHRR	Aryl-hydrocarbon receptor repressor	P185A	Rs2292596	0.122
AKR1C3	Aldo-keto reductase family 1, member C3	H5Q	Rs12529	0.122
APOB	Apoliprotein B	N338S	Rs1042034	0.177
APOB	Apoliprotein B	P2739L	Rs676210	0.177
ATM	Ataxia telangiectasia mutated	E126D	Rs2234997	0.115
BRCA1	Breast cancer 1, early onset	P871L	Rs79917	0.172
CASR	Calcium-sensing receptor	R990G	Rs1042636	0.121
CD86	CD86 antigen	I179V	Rs2681417	0.162
CDKN1B	Cyclin-dependent kinase inhibitor 1B	V109G	Rs2066827	0.223
CYP1B1	Cytochome P450, subfamily 1	L432V	Rs1056836	0.346
CYP2D6	Cytochrome p450, subfamily IID	C265Y	Rs2743458	0.209
DHDH	Dihydrodiol dehydrogenase	S66N	Rs2270941	0.126
EPHX1	Epoxide hydrolase, microsomal	Y133H	Rs1051740	0.152
GSTM3	Glutathione *S*-transferase M3	V224I	Rs7483	0.160
IL4R	Interlekin-4 receptor	Q576R	Rs10801275	0.226
LEPR	Leptin receptor	K109R	Rs1137100	0.137
NAT2	*N*-Acetyl transferase 2	K268R	Rs1208	0.116
SHMT1	Serine hydroxymethyl transferase 1	L474F	Rs1979277	0.162
SLC23A1	Solute carrier family 23, member 1	V264M	—	0.158
TNFRSF10A	Tumor necrosis factor receptor superfamily, member 10A	T209R	—	0.260

[a] Single-letter amino acid code is used; amino acid positions are numbered as in the primary translation product.
[b] Identification (ID) number in the public SNP database, dbSNP (Sherry *et al.* 2001).

chosen because of their potential involvement in cancer or other complex diseases (Packer *et al.* 2004). Thus, they included loci encoding proteins involved in fundamental cellular processes, many of which are expected to be subject to strong purifying selection because of stringent functional constraints. On the other hand, they also included a large number of loci encoding cytokines and other immune system proteins, which are known to evolve rapidly at the amino acid level and thus appear to be subject to a relatively low level of functional constraint (Murphy 1993; Hughes 1997). Further studies will be required to determine whether the frequency of sites subject to purifying selection in our data set differs from that in a random set of human protein-coding loci.

The existence of slightly deleterious alleles subject to purifying selection yet nonetheless occurring at rather high frequencies in the human population is consistent with Ohta's (1976, 2002) nearly neutral theory of molecular evolution. This theory predicts that, when the effective population size is small, slightly deleterious alleles behave as if selectively neutral and thus can drift to higher than expected frequencies in large populations (Ohta 1976). There is substantial evidence that the human population has undergone an expansion beginning in the Pleistocene, and thus that our species has a bottlenecked population history (Harpending *et al.* 1998; Marth *et al.* 2003). Slightly deleterious alleles that drifted to high frequencies during the bottleneck phase are expected to decrease gradually in frequency as effective population size increases and purifying selection becomes more effective. SNPs in the 5′-UTR and radical nonsynonymous SNPs showed a reduction in gene diversity in comparison with other SNPs from the same set of genes that is consistent with this prediction.

By contrast, a number of SNPs in the 5′-UTR and a number of nonsynonymous SNPs showed very high genetic distances between African and Asian populations (Table 4). Although these high values might be due to genetic drift or stochastic error, it is possible that at least some of these represent cases where positive selection has acted to favor different phenotypes in different human populations. Interestingly, these cases include immune system genes (IL8 and LIG4), hormones and hormone receptors (GH1 and GNRH1), and metabolic enzymes (*e.g.*, ADH1B and CYP1B1), which might plausibly be subject to different selective pressures in different environments.

Published studies on several SNPs identified in our study provide an important context for assessing the utility of this strategy. For example, a well-characterized SNP (rs4073), which lies in the proximal promoter of IL8, has been reported to alter the regulation of the IL8 gene and has been associated with two major infectious disease outcomes, resistance to tuberculosis and severity of respiratory syncytial virus (RSV) infection (Hull *et al.* 2001; Choi *et al.* 2002; Ma *et al.* 2003). Interestingly, the association between IL8 and RSV has been shown in Europeans but not in Koreans, where the distribution of variant alleles is markedly shifted. Similarly, IL4R has been associated with RSV infection in European children but not in Korean children—again suggesting that variants in IL8 or IL4R, or perhaps in additional genes, may contribute to the observed differences between populations even for SNPs in which a functional difference between alleles has been demonstrated (Hoebee *et al.* 2003). Several other notable examples include studies by Rockman *et al.* (2003) in which they provide evidence that natural selection has led to different allelic frequencies in human populations at a SNP locus in the promoter of another cytokine, IL4, and Bernig *et al.* (2004) in which they show different frequencies of functionally significant haplotypes for the MBL2 gene.

A metabolic gene exhibiting SNPs with high interpopulation differences is the ADH1B gene, which exhibits both radical (R370C) and conservative (R48H) nonsynonymous SNPs with high interpopulation divergence. In particular, Osier *et al.* (2002) considered the frequency divergence between East Asian and other human populations at the R48H SNP, which exhibits the highest F_{st} observed among 86 unlinked sites in 33 worldwide populations (Osier *et al.* 2002), too high to be explainable by genetic drift alone. A group of metabolic genes exhibiting SNPs with high interpopulation differences are those involved in oxidation of benzo[α]pyrene, including CYP1B1, DHDH, and EPHX1, each exhibiting a conservative nonsynonymous SNP with a significant interpopulation difference (Table 4). In particular, CYP1B1, with a SNP (L432V) with the highest interpopulation difference in the samples examined in this study, has SNPs with well-characterized population differences, *e.g.*, a *d* of 0.223 in a sample of 111 and 120 individuals with African and Asian ancestry, respectively (Mammen *et al.* 2003).

The hypothesis of positive selection at the SNP sites listed in Table 4 can be tested further by examining allelic frequencies in larger samples and also by testing for associations between polymorphism at these loci and observable phenotypes. Thus, the computation of genetic distances between populations at SNP sites provides a simple strategy for identifying SNPs that may have been subject to directional selection over the course of human evolution, which can complement strategies based on statistics such as F_{ST} (Akey *et al.* 2002) and linkage disequilibrium (Schneider *et al.* 2003) or statistics based on the pattern of sequence variability (Bamshad and Wooding 2003).

This project has been funded in part with federal funds from the National Cancer Institute, National Institutes of Health (NIH), under contract no. NO1-CO-12400. Partial support was provided by NIH grant GM43940 to A.L.H.

LITERATURE CITED

Akey, J. M., G. Zhang, K. Zhang, L. Jin and M. D. Shriver, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Res. **12:** 1805–1814.

Bamshad, M., and S. P. Wooding, 2003 Signatures of natural selection in the human genome. Nat. Rev. Genet. **4:** 89–111.

Bernig, T., J. G. Taylor, B. Staats, C. B. Foster and M. Yeager, 2004 Sequence analysis of the mannose binding lectin (MBL2) gene reveals a high degree of heterozygosity with evidence of selection. Genes Immun. **5:** 461–476.

Botstein, D., and N. Risch, 2003 Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nat. Genet. **33** (Suppl.): 228–237.

Brookes, A. J., 1999 The essence of SNPs. Gene **234:** 177–186.

Chanock, S., 2001 Candidate genes and single nucleotide polymorphisms (SNPs) in the study of human disease. Dis. Markers **19:** 89–98.

Choi, E., H. J. Lee, T. W. Yoo and S. J. Chanock, 2002 A common haplotype of interleukin 4 gene is associated with severe respiratory syncytial virus disease in Korean children. J. Infect. Dis. **186:** 1207–1211.

Fleming, M. A., J. D. Potter, C. J. Ramirez, G. K. Ostrander and E. A. Ostrander, 2003 Understanding misssense mutations in the *BRCA1* gene: an evolutionary approach. Proc. Natl. Acad. Sci. USA **100:** 1151–1156.

Freudenberg-Hua, Y., J. Freudenberg, N. Kluck, S. Cichon, P. Propping *et al.*, 2003 Single nucleotide variation analysis in 65 candidate genes for CNS disorders in a representative sample of the European population. Genome Res. **13:** 2271–2276.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Gray, I. C., D. A. Campbell and N. G. Spurr, 2000 Single nucleotide polymorphisms as tools in human genetics. Human Mol. Genet. **9:** 2403–2408.

Hammer, M. F., F. Blackmer, D. Garrigan, M. W. Nachman and J. A. Wilder, 2003 Human population structure and its effect on sampling Y chromosome sequence variation. Genetics **164:** 1495–1509.

Harpending, H. C., M. A. Batzer, M. Gurven, L. B. Jorde, A. R. Rogers *et al.*, 1998 Genetic traces of ancient demography. Proc. Natl. Acad. Sci. USA **95:** 1961–1967.

Hoebee, B., E. Rietveld, L. Bont, M. Oosten, H. M. Hodemaekers *et al.*, 2003 Association of severe respiratory syncytial virus bronchiolitis with interleukin-4 and interleukin-4 receptor alpha polymorphisms. J. Infect. Dis. **187:** 2–11.

Hughes, A. L., 1997 Rapid evolution of immunoglobulin C2 domains expressed in immune system cells. Mol. Biol. Evol. **14:** 1–5.

Hughes, A. L., B. Packer, R. Welch, A. W. Bergen, S. J. Chanock *et al.*, 2003 Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc. Natl. Acad. Sci. USA **100:** 15754–15757.

Hull, J., H. Ackerman, K. Isles, S. Usen, M. Pinder *et al.*, 2001 Unusual haplotypic structure of IL8, a susceptibility locus for a common respiratory virus. Am. J. Hum. Genet. **69:** 413–419.

Kimura, M., and T. Ohta, 1974 On some principles governing molecular evolution. Proc. Natl. Acad. Sci. USA **71:** 2848–2852.

Lohnmueller, K. E., C. L. Pearce, M. Pike, E. S. Lander and J. N. Hirschhorn, 2003 Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat. Genet. **33:** 177–182.

Ma, X., R. A. Reich, J. A. Wright, H. R. Tooker, L. D. Teeter *et al.*, 2003 Association between interleukin-8 gene alleles and human susceptibility to tuberculosis disease. J. Infect. Dis. **188:** 349–355.

Mammen, J. S., G. S. Pittman, Y. Li, F. Abou-Zahr, B. A. Bejjani *et al.*, 2003 Single amino acid mutations, but not common polymorphisms, decrease the activity of CYP1B1 against (-)benzo[a]-pyrene-7R-trans-7,8-dihydrodiol. Carcinogenesis **24:** 1247–1255.

Marth, G., G. Shuler, R. Yeh, R. Davenport, R. Agarwala *et al.*, 2003 Sequence variations in the public human genome data reflect a bottlenecked population history. Proc. Natl. Acad. Sci. USA **100:** 376–381.

McKusick, V. A., and C. A. Francomano, 1997 *Mendelian Inheritance in Man: A Catalog of Human Genes and Genetic Disorders*, Ed. 12. Johns Hopkins University Press, Baltimore.

Miyata, T., S. Miyazawa and T. Yasunaga, 1979 Two types of amino acid substitution in protein evolution. J. Mol. Evol. **12:** 219–236.

Murphy, P. M., 1993 Molecular mimicry and the generation of host defense protein diversity. Cell **72:** 823–826.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei, M., and G. Livshits, 1989 Genetic relationships of Europeans, Asians and Africans and the origin of modern *Homo sapiens*. Hum. Hered. **39:** 276–281.

Ng, P. C., and S. Henikoff, 2002 Accounting for human polymorphisms predicted to affect protein function. Genome Res. **12:** 436–446.

Ohta, T., 1976 Role of very slightly deleterious mutations in molecular evolution and polymorphism. Theor. Popul. Biol. **10:** 254–275.

Ohta, T., 2002 Near-neutrality in evolution of genes and gene regulation. Proc. Natl. Acad Sci USA **99:** 16134–16137.

Osier, M. V., A. J. Pakstis, H. Soodyall, D. Comas, D. Goldman *et al.*, 2002 A global perspective on genetic variation at the ADH gene reveals unusual patterns of linkage disequilibrium. Am. J. Hum. Genet. **71:** 84–99.

Packer, B., M. Yeager, B. Staats, R. Welch, A. Crenshaw *et al.*, 2004 SNP500Cancer: a public resource for sequence validation and assay development for genetic variation in candidate genes. Nucleic Acids Res. **32:** D528–D532.

Ramensky, V., P. Bork and S. Sunyaev, 2002 Human non-synonymous SNPs: server and survey. Nucleic Acids Res. **30:** 3894–3900.

Risch, N. J., 2000 Searching for genetic determinants for the new millennium. Nature **405:** 847–856.

Rockman, M. V., M. W. Hahn, N. Soranzo, D. B. Goldstein and G. A. Wray, 2003 Positive selection on a human-specific transcription factor binding site regulating IL4 expression. Curr. Biol. **13:** 2118–2123.

Schneider, J. A., M. S. Pungliya, J. Y. Choi, R. Jiang, X. J. Sun *et al.*, 2003 DNA variability of human genes. Mech. Ageing Dev. **124:** 17–25.

Schork, N. J., D. Fallin and S. Lanchbury, 2000 Single nucleotide polymorphisms and the future of genetic epidemiology. Clin. Genet. **58:** 250–264.

Sherry, S. T., M.-H. Warc, M. Kholodov, J. Baler, E. M. Smigielski *et al.*, 2001 dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. **29:** 308–311.

Stitzel, N. O., Y. Y. Tseng, D. Pervouchine, D. Goddeau, S. Kasif *et al.*, 2003 Structural location of disease-associated single-nucleotide polymorphisms. J. Mol. Biol. **327:** 1021–1030.

Sunyaev, S., V. Ramensky, I. Koch, W. Lathe, III, A. S. Kondrashov *et al.*, 2001 Prediction of deleterious human alleles. Hum. Mol. Genet. **10:** 591–597.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Tishkoff, S. A., and B. C. Verrelli, 2003 Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu. Rev. Hum. Genet. **4:** 293–340.

Wang, Z., and J. Moult, 2001 SNPs, protein structure, and disease. Hum. Mutat. **17:** 263–270.

Wong, G. K.-S., Z. Yang, D. A. Passey, M. Kibukawa, M. Paddock *et al.*, 2003 A population threshold for functional polymorphisms. Genome Res. **13:** 1873–1879.

Zhao, Z., Y.-X. Fu, D. Hewett-Emmett and E. Boerwinkle, 2003 Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene **312:** 207–213.

Communicating editor: S. Schaeffer